Proceedings of Machine Learning Research vol 75:1–21, 2018 31st Annual Conference on Learning Theory

Incentivizing Exploration by Heterogeneous Users

Bangrui Chen BC496@CORNELL.EDU
Peter I. Frazier PF98@CORNELL.EDU
Operations Research and Information Engineering
Cornell University
New York, NY 14850, USA

David Kempe DAVID.M.KEMPE@GMAIL.COM
Department of Computer Science
University of Southern California
Los Angeles, CA 90089, USA

Editors: Sébastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We consider the problem of incentivizing exploration with heterogeneous agents. In this problem, N bandit arms provide vector-valued outcomes equal to an unknown arm-specific attribute vector, perturbed by independent noise. Agents arrive sequentially and choose arms to pull based on their own private and heterogeneous linear utility functions over attributes and the estimates of the arms' attribute vectors derived from observations of other agents' past pulls. Agents are myopic and selfish and thus would choose the arm with maximum estimated utility. A principal, knowing only the distribution from which agents' preferences are drawn, but not the specific draws, can offer arm-specific incentive payments to encourage agents to explore underplayed arms. The principal seeks to minimize the total expected cumulative regret incurred by agents relative to their best arms, while also making a small expected cumulative payment.

We propose an algorithm that incentivizes arms played infrequently in the past whose probability of being played in the next round would be small without incentives. Under the assumption that each arm is preferred by at least a fraction $p > 0$ of agents, we show that this algorithm achieves expected cumulative regret of $O(N\mathrm{e}^{2/p} + N\log^3(T))$, using expected cumulative payments of $O(N^2\mathrm{e}^{2/p})$. If p is known or the distribution over agent preferences is discrete, the exponential term $\mathrm{e}^{2/p}$ can be replaced with suitable polynomials in N and $1/p$. For discrete preferences, the regret's dependence on T can be eliminated entirely, giving constant (depending only polynomially on N and $1/p$) expected regret and payments. This constant regret stands in contrast to the $\Theta(\log(T))$ dependence of regret in standard multi-armed bandit problems. It arises because even unobserved heterogeneity in agent preferences causes exploitation of arms to also explore arms fully; succinctly, heterogeneity provides free exploration.

Keywords: Incentivizing Exploration, Multi-Armed Bandits, Social Learning

1. Introduction

Many websites and apps facilitate joint discovery, sharing, and recommendation of content. This includes news-, photo-, and video-sharing sites; sites that host user-written reviews of products, restaurants, hotels, and travel experiences; and citizen science projects such as eBird (Sullivan et al., 2009; Xue et al., 2013) and Galaxy Zoo (Lintott et al., 2008). Users of these sites improve their own experiences by learning from the experiences of others (Schmit and Riquelme, 2018).

Viewed more abstractly, users jointly explore a space of options (products, news stories, photos, birdwatching sites, . . .), with the implicit goal of identifying the "best" ones. Such scenarios are often modeled in a bandit learning framework. However, unlike standard bandit settings, the utilities of the decision makers (the users) are not aligned with overall utility. Societally (i.e., in a suitable aggregate over users), substantial exploration of options could provide higher future rewards. However, individual users interact with the site a limited number of times, and therefore have little incentive to explore. Recent work has focused on the setting in which users interact only once, and therefore have no intrinsic incentive to explore.

Two recent lines of work have shown that effecting a societally near-optimal outcome in this setting requires explicitly inducing exploration: Kremer et al. (2014) and Mansour et al. (2015, 2016, 2018) (see Slivkins (2017) for an overview) assume that the site (also called the *principal*) has an informational advantage in being the only one to observe the results of past arm pulls (as in driving route recommendations). The principal can use her[1] advantage to induce exploration by recommending apparently sub-optimal arms, as long as agents cannot do better on their own. Frazier et al. (2014) and Han et al. (2015) instead assume that the results of all past arm pulls are publicly observable (as on a review-sharing site). They suppose that the principal can incentivize exploration by offering arm-specific reward payments. We follow this second model.

Past work has assumed that users have homogeneous preferences over arms, i.e., the expected reward derived from an arm is the same for all users. (Han et al. (2015) model users that are heterogeneous in their tradeoff between utility derived from arm pulls and utility derived from the principal's payment, but assume that preferences across arms are homogeneous.) In reality, users have different preferences, e.g., gastronomic, political, aesthetic, practical. Indeed, the websites and mobile apps most widely used for joint discovery, sharing, and recommendation of content tend to concern products and items with heterogeneity in preferences (movies, restaurants, videos, travel experiences), and not items with a universally agreed-on best order. This is perhaps because regimes with heterogeneous preferences are the ones where people have the most difficulty discovering the best items, and thus where online platforms provide the most value. Thus, we see an appropriate accounting for heterogeneity as critical to incentivizing exploration in online communities.

In this work, we present the first algorithm and analysis (of which we are aware) for incentivizing exploration when users have heterogeneous preferences over arms. Our analysis provides insight into the impact of user heterogeneity on the principal's ability to achieve high social utility with low incentive payments, and on the best approaches for doing so. Heterogeneity presents both a challenge and an opportunity. On the one hand, unobserved heterogeneity hides critical information about an agent's preferences from the principal, making her unaware of agents' preferred arms. Thus, even with an unlimited incentive budget, incentivizing pulls according to a standard multi-armed bandit algorithm does not provide low regret. Instead, she must use incentives sparingly and allow agents to reveal their preferences through action. On the other hand, heterogeneity also offers the possibility of "free exploration." Even unincentivized, agents will play a variety of arms, revealing information about their attributes. This stands in sharp contrast to the case of homogeneous preferences, where unincentivized agents will herd onto a single apparently best arm, and where effecting essentially any exploration at all requires incentives.

We describe our model at a high level here, with formal definitions given in Section 2. The N arms and users (or *agents*) are characterized by payoff-relevant *attribute* (or *feature*) vectors. Arms'

1. We use male pronouns to refer to users and female pronouns to refer to the principal.

attributes are a priori unknown, and agents' attributes are drawn from a known distribution. An agent's reward from pulling an arm is the inner product of his vector with the arm's vector (plus noise). When an arm is pulled, a noisy version of its attribute vector is observed by everyone. This models full-text product reviews on websites like Amazon, or ratings of "service", "value", and other restaurant attributes on websites like Tripadvisor. Agents are myopic and will pull the arm whose expected attribute vector (based on past noisy observations) maximizes their reward. This assumption requires agents to have access to aggregate information about the feature vector estimates of all (relevant) arms. Such access is facilitated by the platforms' ability to aggregate and effectively display past feedback, e.g., by displaying simple averages of ratings and using automatic summarization of full-text reviews (Wang and Yaman, 2010; Liu et al., 2012; Abulaish et al., 2009). While platforms could in principle manipulate the display of such data, most have an incentive to be seen as truthful. The principal can incentivize agents to pull particular arms by offering arm-specific payments. The principal's goal is to keep the cumulative regret across all agents small, while incurring only small total payments.

Our main theorem can be stated informally as follows:

Theorem 1 *Assume that for each arm, at least a constant fraction p of the population likes this arm best, and let L be a measure of the "density of near-ties" between agents' arm preferences (in a sense made precise in Section 2). There is a policy that achieves expected cumulative regret $O(N\mathrm{e}^{2/p} + LN\log^3(T))$, using expected cumulative payments of $O(N^2\mathrm{e}^{2/p})$. In particular, when agents who are close to tied between two arms have measure 0, both the expected regret and expected payment are bounded by constants (with respect to T).*

The policy achieving the result of Theorem 1 is simple: it mostly lets agents exploit arms, but incentivizes them to explore when arms appear unlikely to be pulled without incentives. It is presented in detail in Section 4.

2. Preliminaries

We consider a multi-armed bandit setting with N arms. Arm payoffs are determined by d *attributes* or *features*; hence, arms are identified with vectors $\boldsymbol{\mu}_i \in \mathbb{R}^d$. The $\boldsymbol{\mu}_i$ are (adversarially) fixed, and unknown to the agents and the principal. When arm i is pulled, its current utility-relevant features are determined as $\boldsymbol{\mu}_i + \boldsymbol{\zeta}$, where $\boldsymbol{\zeta}$ is a mean-zero independent continuous sub-Gaussian[2] noise vector $\boldsymbol{\zeta} \sim \mathrm{subG}(\mathbf{0}, \sigma^2 I_d)$. Here, I_d denotes the $d \times d$ identity matrix.

At each time t, a new user (or *agent*) arrives, whose feature vector (which we also call his *type*) $\boldsymbol{\theta}_t \in \mathbb{R}^d$ is drawn from a known distribution f. Depending on context, we will identify agents with their arrival time t or their type $\boldsymbol{\theta}_t$. When agent t pulls arm $i_t := i$, he and all future agents observe a vector $\boldsymbol{y}_t = \boldsymbol{\mu}_i + \boldsymbol{\zeta}_t$ for arm i, and his reward is $\boldsymbol{\theta}_t \cdot \boldsymbol{y}_t$, i.e., agents have linear preferences. Although the model is linear, it permits additional flexibility through the addition of derived attributes that are nonlinear transformation of some original attribute vector. One can simply define $\theta_j = h_j(\boldsymbol{\theta})$ for some known nonlinear functions h_j and $j = d+1, \ldots, d'$, and then increase the number of attributes to d'.

For each time t and arm i, let $m_{t,i}$ be the number of times that arm i has been pulled (strictly) before time t. An agent at time t estimates arm i's attribute vector as the average of vectors observed

2. For simplicity of notation, we assume that the variance proxy σ^2 is uniform across time steps. The analysis extends easily when the variance proxy changes over time.

during the arm's past pulls: $\hat{\boldsymbol{\mu}}_{t,i} = \frac{1}{m_{t,i}+1} \cdot (\hat{\boldsymbol{\mu}}_{0,i} + \sum_{t'<t:i_{t'}=i} \boldsymbol{y}_{t'})$; here, $\hat{\boldsymbol{\mu}}_{0,i}$ is a single draw $\boldsymbol{\mu}_i + \boldsymbol{\zeta}$ for arm i. (In other words, we assume each arm is pulled once for free at time 0.) For a justification of the assumption that agents can observe the actual noisy vectors $\boldsymbol{y}_{t'}$ of past pulls, see Section 1.

Since each user only pulls an arm once, users are *myopic*: in the absence of incentives, user t will pull an arm from $\operatorname{argmax}_i \boldsymbol{\theta}_t \cdot \hat{\boldsymbol{\mu}}_{t,i}$. To incentivize users to explore more, the principal can offer *payments* $c_{t,i}$ to user t for pulling arm i. Then, user t will pull an[3] arm i maximizing $(c_{t,i} + \boldsymbol{\theta}_t \cdot \hat{\boldsymbol{\mu}}_{t,i})$. The principal cannot observe $\boldsymbol{\theta}_t$, and only knows the distribution f from which it is drawn. Her goal is to reduce both the cumulative *regret* experienced by all users up to time T, and the total *payment* she makes to the users.

We define the regret at time t as $r_t = (\max_i \boldsymbol{\theta}_t \cdot \boldsymbol{\mu}_i) - \boldsymbol{\theta}_t \cdot \boldsymbol{\mu}_{i_t}$, and the cumulative regret up to time T as $R_T = \sum_{t=1}^{T} r_t$. Similarly, $c_t = c_{t,i_t}$ is the actual incentive payment at time t, and the cumulative payment up to time T is $C_T = \sum_{t=1}^{T} c_t$. More formally, the principal's goal is to find a policy $\mathcal{A}$ for offering payments under which both the cumulative expected regret $\mathbb{E}[R_T]$ the cumulative expected payment $\mathbb{E}[C_T]$ are small.

We assume above that the distribution f over preference vectors is known to the principal. In practice, a principal would estimate this distribution from agents' selections and the attribute estimates and offered payments on which they were based. While we do not show it, we hypothesize that our regret guarantees only change by constants as long as the principal's estimate of the conditional probability of pulling an arm is correct to within a constant factor.

To support the formulation of our results and the analysis, we define the following additional notation. We let $B_{\boldsymbol{\theta}} \in \operatorname{argmax}_i \boldsymbol{\theta} \cdot \boldsymbol{\mu}_i$ and $B'_{\boldsymbol{\theta}} \in \operatorname{argmax}_{i \neq B_{\boldsymbol{\theta}}} \boldsymbol{\theta} \cdot \boldsymbol{\mu}_i$ denote the (indices of) the best and second-best arms for an agent with attribute vector $\boldsymbol{\theta}$, breaking ties arbitrarily (but consistently). Notice that based on Assumption 3 below, $B_{\boldsymbol{\theta}}$ is unique with probability 1.

Our algorithms rely on the following three assumptions on the agent distribution f.

Assumption 1 (Compact Support) *f has a compact support set contained in $[0, D]^d$.*

Let $p = \min_i \operatorname{Prob}_{\boldsymbol{\theta} \sim f}[B_{\boldsymbol{\theta}} = i]$ denote the minimum (over all arms) fraction of users that prefer any particular arm.

Assumption 2 (Every arm is someone's best) *Each arm i has a strictly positive proportion of users for whom i is the best arm; that is, $p > 0$.*

Let $q(z)$ be the cumulative distribution function (CDF) $q(z) = \operatorname{Prob}_{\boldsymbol{\theta} \sim f}\left[(\boldsymbol{\mu}_{B_{\boldsymbol{\theta}}} - \boldsymbol{\mu}_{B'_{\boldsymbol{\theta}}}) \cdot \boldsymbol{\theta} \leq z\right]$. In words, $q(z)$ is the CDF of the *strength* of the preference of a random agent for his best arm over his second-best arm.

Assumption 3 (Not too many near-ties) *Near-ties have vanishing probability; that is, there exist constants $\hat{z} > 0, L$ such that $q(z) \leq L \cdot z$ for all $z \leq \hat{z}$.*

Assumption 3 implies that ties happen with probability 0. One special case of interest we discuss below is $L = 0$. This case arises when there is a cutoff $\hat{z}$ such that only a measure-zero set of agents has two best arms within utility $\hat{z}$ of each other. In particular, $L = 0$ happens whenever f is supported on a finite set of types, and each type has a unique best arm.

3. We assume that ties are broken in favor of an arm with largest payment $c_{t,i}$.

Roles of parameters: Our problem setting is characterized by many parameters. Of these, we consider N, $p \leq 1/N$ and T to be the key parameters, and also give some prominence to L to illustrate an interesting phenomenon. On the other hand, we consider σ, D, d, and $\hat{z}$ to be constant. We track all parameters through most of our proofs, but report final results in terms of only the key parameters, except where we illustrate a specific point. In particular, we consider d a constant, because users typically evaluate products (or restaurants, etc.) by a small number of relevant features.

3. Overview of Results and Discussion

Our main algorithm is presented as Algorithm 1 in Section 4. Our main result is the following pair of theorems[4], analyzing the payments and regret of the algorithm.

Theorem 2 *The expected total payment of Algorithm 1 is at most* $O\left(N^2 \cdot \mathrm{e}^{2/p}\right)$.

Theorem 3 *For any time horizon T, the expected cumulative regret for Algorithm 1 up to time T is bounded above by* $O\left(N \cdot \mathrm{e}^{2/p} + LN \log^3(T)\right)$.

When $L = 0$, the bound of Theorem 3 is constant in T; thus, the algorithm achieves constant regret using constant expected payments. As discussed in Section 2, the case $L = 0$ arises, for instance, for discrete agent distributions with finite support. In fact, when $L = 0$, one can also modify the algorithm to reduce the dependence on p from exponential to polynomial. Theorem 9 (given later) states that the modified algorithm achieves expected regret $O\left(N/p\right)$ with expected payments of $O\left(N^2/p\right)$.

The fact that constant regret can be achieved with constant payment (independent of T) when $L = 0$ suggests aiming for a constant bound more generally, i.e., for $L > 0$. That such a bound is unachievable is shown in Appendix A, where we show a lower bound of $\Omega(\log(T))$ on the expected regret of any algorithm. The instance is simple: it has two arms, one with known attributes; in addition, one draw from the other arm is observed in each step t even when it is not pulled. While the probability of pulling the wrong arm decreases over time, it does not do so fast enough, causing the stated regret.

The exponential dependence on $1/p$ implies an exponential dependence on N (because $p \leq 1/N$). This exponential dependence arises from a need to continue to incentivize arm pulls to ensure that nearly tied agents learn their best arms quickly. Aside from the assumption that $L = 0$, another assumption allows us to eliminate this exponential dependence. Namely, when p (or a lower bound on it) is known ahead of time, the algorithm can be modified to incentivize arms less aggressively. As shown in Theorem 10, the modified algorithm has expected regret at most $O\left(\frac{N}{p^3} + \frac{NL\log^3(T)}{p}\right)$, with expected payments of at most $O\left(\frac{N^2}{p^{5/2}}\right)$.

We compare these bounds to those for standard bandits, focusing on the dependence on T. The standard bandit setting is the case when the agent types $\boldsymbol{\theta}$ are concentrated on a single point, and agents pull arms at the principal's direction without requiring payment. (This setting violates our Assumption 2, so our bounds do not apply to it.) Then, the payment is 0 and the expected regret scales as $\Theta(\log(T))$ (Bubeck and Cesa-Bianchi, 2012, Theorem 2.1).

Our algorithm's payment is constant in T, while its regret is $O(\log^3(T))$ in general with a lower bound of $\Omega(\log(T))$; when preferences are discrete, our algorithm's regret is constant in T. Thus,

4. Recall that we omit the dependence on parameters other than N, p, T unless making a particular point.

viewed solely in terms of the dependence on T, the best performance achievable seems comparable to that in a standard multi-armed bandit problem; but when preferences are discrete, the constant regret surpasses the $\Theta(\log(T))$ achievable in the standard multi-armed bandit setting. This may seem surprising, because the principal in our setting has both less control and less information than in the standard bandit setting. The result arises because heterogeneity in preferences provides free exploration, and allows all of the arms to be pulled infinitely often without incurring regret once estimates are accurate enough.

While heterogeneity in preferences enables this free exploration, heterogeneity alone is not always sufficient for enabling performance improvements compared to the standard bandit setting. Indeed, suppose that agents are still heterogeneous, but the principal pulls arms directly. Unless the principal can also observe the agents' types, she will be unable to correctly choose each agent's preferred arm, even with infinite exploration of arm attributes. Regret will then grow as $\Omega(T)$.

Thus, reaping the benefits of (unobserved) heterogeneous preferences requires the principal to give up direct control of the arms, providing agents the autonomy they need to express their private information about their own preferences. Our results show that simple arm-based incentives are sufficient to overcome the apparent challenges created by this abdication of control.

4. Main Algorithm and Analysis

The algorithm achieving the bounds of Theorems 2 and 3 is simple. It mostly allows agents to exploit, but when an arm is sufficiently unlikely to be pulled, it incentivizes this arm with a payment high enough to guarantee that the next agent pulls it. This way, the algorithm ensures that each arm is pulled often enough.

More precisely, the algorithm divides time into *phases* $s = 1, 2, 3, \ldots$. Phase s starts when each arm has been pulled at least s times. We indicate the start time of phase s by t_s. An arm i is *payment-eligible* at time t (in phase s) if both of the following hold:

- i has been pulled at most[5] s times up to time t, i.e., $m_{t,i} \leq s$.
- The conditional probability of pulling arm i is less than $1/\log(s)$ given the current estimates $\hat{\boldsymbol{\mu}}_{t,i'}$ of the arms' attribute vectors. In other words, we require $\phi_{t,i} < 1/\log(s)$ where $\phi_{t,i} = \text{Prob}_{\boldsymbol{\theta}\sim f}\left[\boldsymbol{\theta}\cdot\hat{\boldsymbol{\mu}}_{t,i} > \boldsymbol{\theta}\cdot\hat{\boldsymbol{\mu}}_{t,i'} \text{ for all } i'\neq i \mid \hat{\boldsymbol{\mu}}_{t,i'}\ \forall i'\right]$ is the probability that arm i will be pulled by the next (random) agent based on the current estimates. Our assumption of a continuous noise distribution and one free pull ensure that ties in $\boldsymbol{\theta}\cdot\hat{\boldsymbol{\mu}}_{t,i}$ between arms occur with probability 0.

When multiple arms are payment-eligible, the algorithm picks one arbitrarily to incentivize. When the algorithm decides to incentivize an arm i, it offers "whatever it takes," i.e., offers a payment of $c_{t,i} = \max_{\boldsymbol{\theta},i'} \boldsymbol{\theta}\cdot(\hat{\boldsymbol{\mu}}_{t,i'} - \hat{\boldsymbol{\mu}}_{t,i})$. The maximum for $\boldsymbol{\theta}$ is taken over the support of f; recall that we assumed this support to be compact. The payment $c_{t,i}$ may appear unnecessarily high. Indeed, it suffices to incentivize only a $1/\log(s)$ fraction of the agents, and our bounds also hold for an alternate version of our algorithm that offers payment $c_{t,i} = \sup\{c \geq 0 \mid \text{Prob}_{\boldsymbol{\theta}\sim f}\left[c + \boldsymbol{\theta}\cdot\hat{\boldsymbol{\mu}}_{t,i} \geq \max_{i'\neq i}\boldsymbol{\theta}\cdot\hat{\boldsymbol{\mu}}_{t,i'}\right] \leq 1/\log(s)\}$. (This definition ensures that $c_{t,i}$ is well-defined and incentivizes at least a $1/\log(s)$ measure of agents, even if f has discrete points.) However, we focus on the higher payments for simplicity of presentation.

5. in fact: exactly, since the algorithm entered phase s

Notice that $\phi_{t,i}$ depends on the estimates for *all* arms; thus, by pulling another arm i', an arm i may become payment-eligible, or cease to be so. Algorithm 1 gives the full details.

Algorithm 1 Algorithm: Incentivizing Exploration

Set the current phase number $s = 1$. {Each arm is pulled once initially "for free."}
for time steps $t = 1, 2, 3, \ldots$ **do**
  if $m_{t,i} \geq s + 1$ for all arms i **then**
    Increment the phase $s = s + 1$.
  if there is a payment-eligible arm i **then**
    Let i be an arbitrary payment-eligible arm.
    Offer payment $c_{t,i} = \max_{\boldsymbol{\theta},i'} \boldsymbol{\theta} \cdot (\hat{\boldsymbol{\mu}}_{t,i'} - \hat{\boldsymbol{\mu}}_{t,i})$ for pulling arm i (and payment 0 for all other arms).
  else
    Let agent t play myopically, i.e., offer payments 0 for all arms.

The high-level idea in the proofs of our main results, Theorems 2 and 3, is the following. Because the algorithm ensures that each arm is pulled "frequently enough," the estimates $\hat{\boldsymbol{\mu}}_{t,i}$ become gradually more accurate in the phase number s. Thus, the fraction of agents who misidentify their best arm decreases. Because each arm has enough agents that would prefer it based on its true attribute vector, once the arms' attribute vectors are learned well enough, the algorithm will not need to incentivize any more, resulting in a payment bound independent of T. Similarly, instantaneous regret will decrease, and mostly accrue due to "problematic" agents who are nearly tied in their preferences between their top two arms. The detailed analysis following this outline is complicated by dependence between the agents' arm pulls and the estimates which in turn are based on past arm pulls. We begin with several technical lemmas that are used for both the payment and regret bounds.

To formally reason about the event that the estimates of arms' attributes vectors are accurate enough — or fail to be so — we define the events $\mathcal{E}_{t,i,j}(x) := [|\hat{\mu}_{t,i}^{(j)} - \mu_i^{(j)}| \leq x]$ that attribute j of arm i at time t is estimated to within accuracy x or better. Then, $\mathcal{E}_t(x) = \bigcap_{i,j} \mathcal{E}_{t,i,j}(x)$ is the event that at time t, all arm attribute estimates are accurate to within x simultaneously. We will show that for suitable choices of t, x, the events $\mathcal{E}_{t,i,j}(x)$ (and hence, by a union bound, $\mathcal{E}_t(x)$) have high probability, and that when they do, myopic agents do not make large mistakes.

4.1. General Lemmas

We begin by bounding the length of any phase, and more generally, the number of arm pulls by any given set of agents. The bound of Lemma 4 will support bounding the regret of early rounds (before tail bounds have kicked in). The proof of Lemma 4 and all other proofs missing in the main paper may be found in the appendix.

Lemma 4 *For any $s \geq 3$, the expected length of phase s is at most $N \cdot \log(s)$ time steps.*

More generally, for any set of types A, the expected number of times that an agent with a type in A appears in a phase s is at most $f(A) \cdot N \cdot \log(s)$, where $f(A) := \text{Prob}_{\boldsymbol{\theta} \sim f}[\boldsymbol{\theta} \in A]$.

We now state the key technical lemma which captures the intuition that the estimates of the arms' attribute vectors become more accurate with increasing phases s.

Lemma 5 *Recall the noise is a mean-zero sub-Gaussian(σ^2) random variable. Let s_0 be fixed, and let $x_n, x'_n > 0$ be sequences satisfying $\sqrt{0.6n \cdot \log(\log_{1.1}(n)+1) + \frac{nx_n^2}{16\sigma^2}} \le \frac{nx'_n}{2\sigma}$, for all $n \ge s_0$. Let τ_s be a stopping time (which may depend on the entire past history) which is almost surely in phase s, i.e., satisfying $\tau_s \in [t_s, t_{s+1})$ almost surely. Then, for any arm i, attribute j, and phase $s \ge s_0$, we have $\text{Prob}\left[\mathcal{E}_{\tau_s,i,j}(x'_s)\right] \ge 1 - 24\exp\left(-\frac{1.8sx_s^2}{16\sigma^2}\right)$.*

Next, we show the complementary result: when the event $\mathcal{E}_t(x)$ happens, no myopic agent incurs large regret.

Lemma 6 *Let $x > 0$ be arbitrary. When $\mathcal{E}_t(x)$ happens, no agent $\boldsymbol{\theta}$ will pull a highly suboptimal arm, i.e., an arm i with $\boldsymbol{\theta} \cdot (\boldsymbol{\mu}_{B_{\boldsymbol{\theta}}} - \boldsymbol{\mu}_i) > 2Ddx$.*

4.2. Bounding the Total Payment

As a first step towards bounding the total payment (and also regret), we show that for sufficiently late phases, under the event $\mathcal{E}_t(x)$ for suitably small x, the algorithm does not offer any payments.

Lemma 7 *Fix an arm i. Let $s \ge \exp(2/p)$, and let τ_s be the (random) time when arm i is pulled for the s^{th} time. Let $\hat{x} = \frac{1}{2Dd} \cdot \min(\hat{z}, \frac{p}{2L})$. Under $\mathcal{E}_{\tau_s}(\hat{x})$, this pull of arm i is not incentivized.*

Towards bounding the algorithm's total payment, we now bound by a constant the *number* of rounds in which the algorithm makes a payment. This bound also turns out to be useful for bounding the total regret.

Lemma 8 *The expected number of time steps in which Algorithm 1 makes any payment is at most $O\left(N\exp\left(\frac{2}{p}\right)\right)$.*

Proof We partition phases into early and late phases. For each of the early phases, we crudely bound the number of payments by N, using that each arm is incentivized at most once per phase. For later phases, we use Lemma 7, which rules out any incentives unless large misestimates of the arm locations occur, which is exponentially unlikely by Lemma 5.

To make this intuition precise, we set $x = x' = \frac{1}{2Dd} \cdot \min(\hat{z}, \frac{p}{2L})$ (which is independent of the phase number s and is equal to $\hat{x}$ defined in Lemma 7). We set the cutoff point between early and late phases to $s_1 = \max(2, \frac{30\sigma^3}{x^3}, \exp(\frac{2}{p}))$. We verify below that $\sqrt{0.6n \cdot \log(\log_{1.1}(n)+1) + \frac{nx^2}{16\sigma^2}} \le \frac{nx}{2\sigma}$ for all $n \ge \max(2, \frac{30\sigma^3}{x^3})$.

Fix an arm i, let $s \ge s_1$, and define τ_s as in Lemma 7. By Lemma 5 (with our choice of $x = x'$), and a union bound over all i, j, we bound the probability $\text{Prob}\left[\mathcal{E}_{\tau_s}(x)\right] \ge 1 - 24Nd \cdot \exp\left(-\frac{1.8sx^2}{16\sigma^2}\right)$. And by Lemma 7, under the event $\mathcal{E}_{\tau_s}(x)$, arm i is not payment-eligible.

Thus, for any $s \ge s_1$, the s^{th} pull of arm i is incentivized with probability at most $24Nd \cdot \exp\left(-\frac{1.8s}{256D^2d^2\sigma^2} \cdot \min(\hat{z}, \frac{p}{L})^2\right)$. Summing over all arms and phases s, adding the at most Ns_1 incentivizations in the first s_1 phases, and using that $\exp(-x) \le 1 - x/2$ for $x \le 1$, the expected

total number of arm incentivizations is at most

$$Ns_1 + 24N^2 d \cdot \frac{1}{1 - \exp\left(-\frac{1.8}{256D^2d^2\sigma^2} \cdot \min(\hat{z}, \frac{p}{L})^2\right)}$$
$$= O\left(\max\left(\frac{NL^3D^3d^3\sigma^3}{p^3}, \frac{ND^3d^3\sigma^3}{\hat{z}^3}, N\exp\left(\frac{2}{p}\right)\right) + \max\left(\frac{N^2d^3L^2\sigma^2D^2}{p^2}, \frac{N^2d^3\sigma^2D^2}{\hat{z}^2}\right)\right)$$
$$= O\left(N\exp\left(\frac{2}{p}\right)\right).$$

It remains to show that $\sqrt{0.6n \cdot \log(\log_{1.1}(n)+1) + \frac{nx^2}{16\sigma^2}} \le \frac{nx}{2\sigma}$ for all $n \ge \max(2, \frac{30\sigma^3}{x^3})$. We first use that $\sqrt{\cdot}$ is sublinear to bound

$$\sqrt{0.6n \cdot \log(\log_{1.1}(n)+1) + \frac{nx^2}{16\sigma^2}} \;\le\; \sqrt{0.6n \cdot \log(\log_{1.1}(n)+1)} + \frac{nx}{4\sigma}$$

(here we use the fact $\sqrt{n} \le n$). Next, we show that $\sqrt{0.6n\log(\log_{1.1}(n)+1)} \le \frac{nx}{4\sigma}$; then, adding the two terms gives the desired bound.

By squaring the claimed statement and rearranging, it is equivalent to show that $\frac{n}{\log(\log_{1.1}(n)+1)} \ge \frac{9.6\sigma^2}{x^2}$. Because $n \ge 2$, a numerical calculation and derivative test shows that $\frac{n}{\log(\log_{1.1}(n)+1)} \ge n^{2/3}$, and because $n \ge \frac{30\sigma^3}{x^3}$, we get that $n^{2/3} \ge \frac{9.6\sigma^2}{x^2}$, completing the proof. ■

It would be desirable to simply identify a constant upper bound on the payment made each time. Unfortunately, while the agent types are drawn from a bounded support, the noise in arm locations is not; hence, with small probability, arm locations may be grossly misestimated, resulting in high incentive payments. As a result, the actual analysis of the total payment is significantly more intricate; the proof of Theorem 2 is given in Appendix F.

4.3. Bounding the Total Regret

In bounding the total regret, because agents' types are from a compact set by Assumption 1, the maximum regret in any one round is bounded by a constant. We use $R = \max_{\boldsymbol{\theta},i,i'} \boldsymbol{\theta} \cdot (\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'})$ to denote this constant upper bound on the maximum regret that can be incurred in any one time step.

Proof of Theorem 3. Regret can arise in two ways: (1) an agent was incentivized to pull a suboptimal arm, or (2) an agent myopically pulled a suboptimal arm. By Lemma 8, Algorithm 1 incentivizes agents at most $O\left(N\exp\left(\frac{2}{p}\right)\right)$ times in expectation, each time causing regret at most R, for a total expected regret of at most $O\left(R \cdot N\exp\left(\frac{2}{p}\right)\right)$. For the rest of the proof, we focus on the regret incurred when agents pull arms myopically and make mistakes.

We distinguish between agents incurring large regret (which requires severe misestimates of arm locations; such misestimates are exponentially unlikely to occur), and agents incurring small positive regret, which requires these agents to be almost tied in their preference for the best arm. To be more precise, we define (with foresight) a phase-dependent cutoff $\gamma(s) = \sqrt{\frac{128\log(s)\cdot D^2d^2\sigma^2}{1.8s}}$, and consider a regret exceeding $\gamma(s)$ large.

For most of the proof, we will focus on the case $\gamma(s) \le \hat{z}$. The remaining phases are the ones with $\frac{s}{\log s} \le \frac{128}{1.8} \cdot D^2d^2\sigma^2/\hat{z}^2$. Because $\frac{s}{\log s} = \Omega(s^{2/3})$, there are at most $O(D^3d^3\sigma^3/\hat{z}^3)$ such

phases, each such phase lasts for at most $N \cdot \log(s)$ steps in expectation by Lemma 4, and each step incurs regret at most R. Therefore, the total expected regret incurred in these phases is at most $O(NR \cdot D^3 d^3 \sigma^3 / \hat{z}^3 \cdot \log(D^3 d^3 \sigma^3 / \hat{z}^3)) \leq O(NRD^4 d^4 \sigma^4 / \hat{z}^4) = O(NR)$, by crudely bounding $\log y \leq O(y^{1/3})$.

For the remainder of the proof, we consider phases s with $\gamma(s) \leq \hat{z}$. We first consider agents $\boldsymbol{\theta}$ incurring positive regret less than $\gamma(s)$. Let A_s be the set of types satisfying $\boldsymbol{\theta} \cdot (\boldsymbol{\mu}_{B_{\boldsymbol{\theta}}} - \boldsymbol{\mu}_{B'_{\boldsymbol{\theta}}}) \leq \gamma(s)$. By Assumption 3, and because $\gamma(s) \leq \hat{z}$, the total measure of A_s is $q(\gamma(s)) \leq L \cdot \gamma(s)$. Then, by Lemma 4, the expected number of pulls in phase s by agents in A_s is bounded above by $L\gamma(s) \cdot N \log(s)$. Any agent $\boldsymbol{\theta}$ incurring positive regret less than $\gamma(s)$ must have a type in A_s, so the expected total regret from such agents, summed over all phases, is at most

$$\sum_{s=1}^{T} L\gamma^2(s) \cdot N \log(s). \tag{1}$$

We next bound the regret incurred in time steps with large regret. Define the stopping times τ_s^k to be the k^{th} time step in the s^{th} phase, with $\tau_s^k = \infty$ when phase s has fewer than k steps, i.e., $k > t_{s+1} - t_s$. By Lemma 6, under $\mathcal{E}_{\tau_s^k}(\frac{\gamma(s)}{2Dd})$, no agent at time t will incur regret more than $\gamma(s)$.

To bound the probability of $\mathcal{E}_{\tau_s^k}(\frac{\gamma(s)}{2Dd})$, we use Lemma 5 (with $x_s = x'_s = \frac{\gamma(s)}{2Dd}$ and a stopping time of τ_s^k). We first verify that this choice of x_s, x'_s satisfies the assumption of Lemma 5 for all $s \geq 2$, i.e., that $\sqrt{0.6n \cdot \log(\log_{1.1}(n) + 1) + \frac{nx_n^2}{16\sigma^2}} \leq \frac{nx_n}{2\sigma}$ for all $n \geq 2$. Substituting that $x_n = \frac{\gamma(n)}{2Dd} = \sqrt{\frac{32 \log(n) \cdot \sigma^2}{1.8n}}$, the left-hand side is

$$\sqrt{0.6n \cdot \log(\log_{1.1}(n) + 1) + \frac{2 \log n}{1.8}} \leq \sqrt{0.6n \cdot \log(\log_{1.1}(n) + 1) + \frac{2n \log n}{1.8}},$$

while the right-hand side is $\sqrt{\frac{8n \log(n)}{1.8}}$. Squaring both sides and canceling out common terms, the desired inequality is equivalent to $\frac{50}{9} \log n \geq \log(\log_{1.1}(n) + 1)$. This can be verified by numerical calculation for $n = 2$ and a derivative test.

Now, by Lemma 5 and a union bound over all arms i and attributes j, we get that

$$\text{Prob}\left[\mathcal{E}_{\tau_s^k}(x_s)\right] \leq 24Nd \cdot \exp\left(\frac{-1.8sx_s^2}{16\sigma^2}\right) = \frac{24Nd}{s^2}.$$

In the low-probability event, we simply bound the maximum regret by R. Summing over all time periods, the total expected regret from large-regret steps is at most $24NdR \cdot \frac{\pi^2}{6} = O(NR)$.

Adding all four types of regret terms, and substituting $\gamma^2(s) = \frac{128 \log(s) \cdot D^2 d^2 \sigma^2}{1.8s}$, the cumulative regret up to time T is at most $O\left(NR \cdot \exp\left(\frac{2}{p}\right) + NR + NR + NL \cdot \sum_{s=1}^{T} \gamma^2(s) \log(s)\right)$ $= O\left(N \cdot \exp\left(\frac{2}{p}\right) + NL \cdot \log^3(T)\right)$. ∎

4.4. Tighter Bounds Under Additional Assumptions

The proofs of Theorem 2 and Theorem 3 incur exponential (in N) payment and regret in the initial phases because the threshold $1/\log(s)$ required for incentivization decreases slowly. This slow

decrease is needed to bound the regret in the later phases when the concentration inequality kicks in as in Equation (1). In this section, we discuss two restrictions under which we can modify the algorithm slightly and provide stronger bounds, avoiding this exponential dependence. Both problem settings are special cases of the more general setting previously considered.

4.4.1. DISCRETE PREFERENCES

Discrete preferences by agents are captured by the following strengthening of Assumption 3, which states that the agents who are close to tied between two arms have measure 0:

Assumption 4 (Discrete Preferences) *There is a positive $\hat{z}$ such that $q(\hat{z}) = 0$.*

When Assumption 4 holds, we restrict the payment-eligibility criterion by only incentivizing arms with much smaller probability to be pulled: an arm i is payment-eligible at time t in phase s when both of the following are true:

- i has been pulled at most s times up to time t, i.e., $m_{t,i} \leq s$.
- Based on the current estimates $\hat{\boldsymbol{\mu}}_{t,i'}$ of all arms' attribute vectors, the probability of pulling arm i is less than $1/s$ (compared to $1/\log(s)$ in the general algorithm).

We refer to this modified version of Algorithm 1 as the *Discrete-Preference Algorithm.* We outline a proof of the following result for this algorithm under Assumption 4 in Appendix G.

Theorem 9 *Under Assumption 4, the Discrete-Preference Algorithm has expected payment bounded above by $O\left(N^2/p\right)$ and expected regret bounded above by $O(N/p)$.*

4.4.2. KNOWN p

An alternative useful assumption is that p (or a strictly positive lower bound on p) is known.

Assumption 5 (Known p) *A strictly positive lower bound on p is known in advance.*

When this assumption holds, we choose a different modification in the definition of payment eligibility. Let $s_0 = \exp(2/p)$. An arm i is *payment-eligible* at time t (in phase s) if both:

- i has been pulled at most s times up to time t, i.e., $m_{t,i} \leq s$.
- Based on the current estimates $\hat{\boldsymbol{\mu}}_{t,i'}$ of all arms' attribute vectors, the probability of pulling arm i is less than $1/\log(s + s_0)$.

This knowledge of p allows the algorithm to incentivize significantly fewer arm pulls. We refer to this modified version of Algorithm 1 as the *Known-p Algorithm.* We outline a proof of the following result for this algorithm under Assumption 5 in Appendix H.

Theorem 10 *Under Assumption 5, the Known-p Algorithm has expected payment bounded above by $O(N^2 \cdot \max(1, (L/p)^{5/2}))$ and expected regret bounded above by $O\left(\frac{N^2}{p^2} + \frac{NL\log^3(T)}{p}\right)$.*

5. Conclusion

We study the problem of incentivizing exploration with heterogeneous user preferences. We proposed an algorithm that achieves expected cumulative regret $O(N\mathrm{e}^{2/p}+N\log^3(T))$, using expected cumulative payments of $O(N^2\mathrm{e}^{2/p})$. It is possible to improve these bounds to polynomial (in N and $1/p$) when p is known or the preference distribution is discrete. In fact, we conjecture that this should be possible even in the full generality of our model. As a first step towards such a polynomial bound, we can obtain an exponential dependence on $1/(pN)$ by changing the probability threshold to be $\frac{1}{N\log(s)}$ [6], which gives polynomial dependence unless some arm has a much smaller fraction of the population preferring it.

Taking this goal one step further, we would like to develop algorithms that do not require all arms to be preferred by a strictly positive fraction of agents. An alternate algorithm might only incentivize an arm if its estimated attribute vector is close enough to a Pareto frontier. The regret will then be $\Omega(\log(T))$ when at least one arm falls below the Pareto frontier, as we no longer have free exploration of all arms. It is likely that a bound will deteriorate as the number of such unpreferred arms increases.

Finally, it would be desirable to generalize to utility functions beyond inner products. We believe that similar results hold for arbitrary Lipschitz-continuous utility functions of the arm's attribute vector, and that only minor modifications are necessary to the algorithm and proofs.

Acknowledgments

PF and BC were partially supported by NSF CMMI-1254298 and AFOSR FA9550-15-1-0038. DK was supported in part by NSF grant 1423618.

References

Muhammad Abulaish, Mohammad Najmud Doja, Tanvir Ahmad, et al. Feature and opinion mining for customer review summarization. In *Proceedings of the 3rd International Conference on Pattern Recognition and Machine Intelligence*, pages 219–224. Springer, 2009.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the 15th ACM conference on Economics and Computation (EC)*, pages 5–22. ACM, 2014.

Li Han, David Kempe, and Ruixin Qiang. Incentivizing exploration with heterogeneous value of money. In *Proceedings of the 11th International Conference on Web and Internet Economics (WINE)*, pages 370–383. Springer, 2015.

Ilan Kremer, Yishay Mansour, and Motty Perry. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012, 2014.

6. This will lead to a different dependence on N in the regret bound as well as the payment bound

Chris J. Lintott, Kevin Schawinski, AnÅe Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg. Galaxy zoo: morphologies derived from visual inspection of ga laxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189, September 2008.

Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou. Movie rating and review summarization in mobile environment. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(3):397–407, 2012.

Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. Bayesian incentive-compatible bandit exploration. In *Proceedings of the 16th ACM Conference on Economics and Computation (EC)*, pages 565–582. ACM, 2015.

Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In *Proceedings of the 17th ACM Conference on Economics and Computation (EC)*, 2016.

Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In *Proceedings of the 9th Innovations in Theoretical Computer Science (ITCS) conference*, 2018.

Sven Schmit and Carlos Riquelme. Human interaction with recommendation systems: On bias and exploration. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Aleksandrs Slivkins. Incentivizing exploration via information asymmetry. *ACM Crossroads*, 24 (1):38–41, 2017.

Brian L. Sullivan, Christopher L. Wood, Marshall J. Iliff, Rick E. Bonney, Daniel Fink, and Steve Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, 2009.

Ye-Yi Wang and Sibel Yaman. Product or service review summarization using attributes, July 1 2010. US Patent App. 12/346,903.

Yexiang Xue, Bistra N. Dilkina, Theodoros Damoulas, Daniel Fink, Carla P. Gomes, and Steve Kelling. Improving your chances: Boosting citizen science discovery. In *Proceedings of the 1st Conference on Human Computation and Crowdsourcing*, 2013.

Shengjia Zhao, Enze Zhou, Ashish Sabharwal, and Stefano Ermon. Adaptive concentration inequalities for sequential decision problems. In *Proceedings of the 30th Advances In Neural Information Processing Systems (NIPS)*, pages 1343–1351, 2016.

Appendix A. A Lower Bound of $\Omega(\log(T))$

We saw in Theorem 9 that when agent preferences for their best arm are sufficiently clear, in the sense that $L = 0$, the regret of Algorithm 1 is bounded by a constant. One may conjecture that

this should hold more generally, in that the regret of the (fewer and fewer) agents on the boundary between close arms should go to 0, while their fraction also goes to 0. In this section, we establish a lower bound, showing that even in very simple settings, a (logarithmic) dependence on T is typically unavoidable for *any* incentivization strategy.

We consider an instance with two arms, whose attribute vectors are $(0,0)$ and $(0,1)$, respectively. Agent types are distributed uniformly on the (edge of the) two-dimensional unit square[7] $\{\boldsymbol{\theta} \mid \max(|\theta_1|,|\theta_2|)=1\}$, with density $\frac{1}{8}$. Because $\boldsymbol{\theta}\cdot(0,0)=0$ and $\boldsymbol{\theta}\cdot(0,1)=\theta_2$, the best choice for agent $\boldsymbol{\theta}$ is arm $(0,1)$ iff $\theta_2>0$, i.e., the top half of the unit square prefers the arm $(0,1)$, and the bottom half prefers the arm $(0,0)$.

Since we are proving a lower bound, we give the algorithm the following extra two advantages: (1) there is no noise in the observations of the arm $(0,0)$, and all agents know its location deterministically. (2) in each time step, regardless of which arm is pulled, the algorithm and all agents observe a pull from arm $(0,1)$. For simplicity of notation, we set the standard deviation of the arm $(0,1)$ to $\sigma=1$; different values only lead to a scaling of the results.

Under these advantages, myopic play is clearly optimal, so it suffices to bound the regret of the myopic algorithm which never incentivizes agents.

Because a pull of arm $(0,1)$ is observed in each time step, after t rounds, the estimate $\hat{\boldsymbol{\mu}}_{t,(0,1)}$ is of the form $(0,1)+(\zeta_{t,1},\zeta_{t,2})$, where $\zeta_{t,1}\sim \mathrm{subG}(0,\sqrt{1/t})$ and $\zeta_{t,2}\sim \mathrm{subG}(0,\sqrt{1/t})$. We lower-bound the regret in step t by focusing on the event that both normal noise coordinates are non-negative, which by symmetry has probability $\frac{1}{4}$:

$$\mathbb{E}\left[r_t\right] \geq \mathbb{E}\left[r_t \mid \zeta_{t,1}>0,\zeta_{t,2}>0\right]\cdot \mathrm{Prob}\left[\zeta_{t,1}>0,\zeta_{t,2}>0\right] \;=\; \frac{1}{4}\mathbb{E}\left[r_t \mid \zeta_{t,1}>0,\zeta_{t,2}>0\right].$$

For the moment, focus on some time step t, and write $\boldsymbol{\zeta}=\boldsymbol{\zeta}_t$. Then, there are two types of agents who pull the wrong arm and incur regret:

1. If $\theta_2>0$ and $\theta_1\zeta_1+\theta_2(1+\zeta_2)<0$ then $\boldsymbol{\theta}$ should pull $(0,1)$, but will wrongly pull $(0,0)$ and incur regret θ_2. The range of $\boldsymbol{\theta}$ making the wrong choice is thus $0<\theta_2<\frac{-\zeta_1}{1+\zeta_2}\cdot\theta_1$. Since we are proving a lower bound, we only focus on the case $\theta_1=-1$, and ignore the case $\theta_2=1$ (which is rare, since it requires ζ_1 to be large). Thus, the set of agents incurring regret contains the set $\{\boldsymbol{\theta}=(-1,\theta_2)\mid 0<\theta_2<\frac{\zeta_1}{1+\zeta_2}\}$.
2. If $\theta_2<0$ and $\theta_1\zeta_1+\theta_2(1+\zeta_2)>0$, then $\boldsymbol{\theta}$ should pull $(0,0)$, but will wrongly pull $(0,1)$ and incur regret $-\theta_2$. This region and its regret are rotationally symmetric to the previous case.

Hence, the expected regret for given ζ_1, ζ_2 is at least $2\int_0^{\frac{\zeta_1}{1+\zeta_2}}\theta_2\cdot\frac{1}{8}\mathrm{d}\theta_2=\frac{1}{8}\cdot\left(\frac{\zeta_1}{1+\zeta_2}\right)^2$. The expected regret, conditioned on $\zeta_1>0$ and $\zeta_2>0$, is therefore

$$\begin{aligned}\mathbb{E}\left[r_t \mid \zeta_1>0,\zeta_2>0\right] &\geq \frac{1}{\mathrm{Prob}\left[\zeta_1>0,\zeta_2>0\right]}\cdot\frac{1}{8}\int_0^\infty\int_0^\infty\left(\frac{\zeta_1}{1+\zeta_2}\right)^2\cdot\frac{\mathrm{e}^{-\frac{t\zeta_1^2}{2}}\sqrt{t}}{\sqrt{2\pi}}\mathrm{d}\zeta_1\frac{\mathrm{e}^{-\frac{t\zeta_2^2}{2}}\sqrt{t}}{\sqrt{2\pi}}\mathrm{d}\zeta_2\\ &=\frac{1}{4\pi}\int_0^\infty\int_0^\infty\left(\frac{\zeta_1}{\sqrt{t}+\zeta_2}\right)^2\mathrm{e}^{-\zeta_1^2/2}\mathrm{e}^{-\zeta_2^2/2}\mathrm{d}\zeta_1\mathrm{d}\zeta_2.\end{aligned}$$

7. While our model technically assumed that all agent coordinates are non-negative, we could simply shift the unit square. The present choice is solely for ease of notation.

We want to show that the expected regret per time step decreases only at a rate of $\Omega(1/t)$, and thereto consider the limit of the ratio of the two quantities:

$$\begin{aligned}\lim_{t\to\infty} \frac{\mathbb{E}\left[r_t \mid \zeta_1 > 0, \zeta_2 > 0\right]}{\frac{1}{t}} &\geq \lim_{t\to\infty} \left(\frac{1}{4\pi} \cdot \int_0^\infty \int_0^\infty t \cdot \left(\frac{\zeta_1}{\sqrt{t}+\zeta_2}\right)^2 \mathrm{e}^{-\zeta_1^2/2} \mathrm{e}^{-\zeta_2^2/2} \mathrm{d}\zeta_1 \mathrm{d}\zeta_2 \right) \\ &= \frac{1}{4\pi} \cdot \int_0^\infty \int_0^\infty \lim_{t\to\infty} \left(t \cdot \left(\frac{\zeta_1}{\sqrt{t}+\zeta_2}\right)^2 \right) \mathrm{e}^{-\zeta_1^2/2} \mathrm{e}^{-\zeta_2^2/2} \mathrm{d}\zeta_1 \mathrm{d}\zeta_2 \\ &= \frac{1}{4\pi} \cdot \int_0^\infty \int_0^\infty \mathrm{e}^{-\zeta_1^2/2} \mathrm{e}^{-\zeta_2^2/2} \mathrm{d}\zeta_1 \mathrm{d}\zeta_2 \;=\; \frac{1}{4\pi};\end{aligned}$$

the second line is due to the monotone convergence theorem, which can be applied because $t\left(\frac{\zeta_1}{\sqrt{t}+\zeta_2}\right)^2$ is strictly increasing in t. Because the expected regret in each time step is at least $\Omega(1/t)$, the total expected regret is at least $\Omega(\sum_{t=1}^{T} \frac{1}{t}) = \Omega(\log(T))$.

Appendix B. Proof of Lemma 4

Lemma 4 *For any $s \geq 3$, the expected length of phase s is at most $N \cdot \log(s)$ time steps.*

More generally, for any set of types A, the expected number of times that an agent with a type in A appears in a phase s is at most $f(A) \cdot N \cdot \log(s)$, where $f(A) := \mathrm{Prob}_{\boldsymbol{\theta}\sim f}\left[\boldsymbol{\theta} \in A\right]$.

Proof We show the second claim for general A, which implies the first claim by taking A to be the support of f.

Fix a phase s, and let S_t be the set of arms that have been pulled at most s times by time t. Let τ_m be the maximum of t_s (the start of phase s) and the time t when S_t first contains (at most) m elements. The phase ends when S_t contains no elements, i.e., $\tau_0 = t_{s+1}$. We consider the expected number of pulls during times $t \in [\tau_m, \tau_{m-1})$ made by agents whose types are in A.

Consider a counter starting from 0 at time τ_m that increments each time an agent arrives with a type in A, and that is stopped the first time an arm in S_t is pulled. For any stopping time $t \in [\tau_m, \tau_{m-1})$, the conditional probability that the counter increments, given the history of pulls, payments, and observations at time t, is $f(A)$. The conditional probability that the counter is stopped by that agent's pull is at least $1/\log(s)$. To see this, consider two cases. In the first case, there is at least one payment-eligible arm, in which case the principal incentivizes an arm in S_t. In the second case, there are no payment-eligible arms. Then, by definition of payment-eligibility, each arm $i \in S_t$ has conditional probability at least $1/\log(s)$ of being pulled.

The expected value of this counter when it is stopped is bounded above by the expected stopped value of another counter whose conditional probability of being stopped is exactly $1/\log(s)$. (This can be shown more formally via a coupling argument.) Let V_ℓ be the conditional expectation of this second counter's value when it is stopped, given that its current value is ℓ and it has not been stopped. Observe that $V_\ell = \ell + V_0$. Enumerating outcomes (either the counter increments or not; either it stops or it continues) gives us the recurrence relation $V_0 = f(A)\frac{1}{\log(s)} + f(A)(1-\frac{1}{\log(s)})V_1 + (1-f(A))(1-\frac{1}{\log(s)})V_0$. Using that $V_1 = 1+V_0$, this equation simplifies to $V_0 = f(A) + (1-\frac{1}{\log(s)})V_0$. Noting that V_0 is finite and solving for V_0 gives us that $V_0 = f(A)\log(s)$. Thus, the conditional

expected number of pulls (given the history at time τ_m) by an agent with a type in A during times in $[\tau_m, \tau_{m-1})$ is no more than $f(A)\log(s)$.

We may write the total number of pulls by agents with types in A during phase s as a sum of this quantity over m ranging from $|S_{t_s}|$ down to 1. Noting that $|S_{t_s}| \leq N$ and taking the expectation of this sum completes the proof. ■

Appendix C. Proof of Lemma 5

Lemma 5 *Recall the noise is a mean-zero sub-Gaussian(σ^2) random variable. Let s_0 be a phase cutoff, and let $x_n, x'_n > 0$ be functions satisfying that $\sqrt{0.6n \cdot \log(\log_{1.1}(n)+1) + \frac{nx_n^2}{16\sigma^2}} \leq \frac{nx'_n}{2\sigma}$, for all $n \geq s_0$. Let τ_s be a stopping time (which may depend on the entire past history) which is almost surely in phase s, i.e., satisfying $\tau \in [t_s, t_{s+1})$ almost surely.*

Then, for any arm i, attribute j, and phase $s \geq s_0$, we have that $\text{Prob}\left[\mathcal{E}_{\tau_s,i,j}(x'_s)\right] \geq 1 - 24\exp\left(-\frac{1.8sx_s^2}{16\sigma^2}\right)$.

The proof of Lemma 5 is based on an adaptive concentration inequality due to Zhao et al. (2016), given as Lemma 11.

Lemma 11 (Corollary 1 of Zhao et al. (2016)) *Let X_i be zero-mean $1/2$-subgaussian random variables, and $\{S_n = \sum_{i=1}^n X_i, n \geq 1\}$ the corresponding random walk. Let J be any stopping time with respect to $\{X_1, X_2, \ldots\}$. (We allow J to take the value ∞, defining* $\text{Prob}\left[J = \infty\right] = 1 - \lim_{n\to\infty}\text{Prob}\left[J \leq n\right]$.*) Define $g(n) = \sqrt{0.6n \cdot \log(\log_{1.1}(n)+1) + n \cdot b}$.*

Then, $\text{Prob}\left[J < \infty \text{ and } S_J \geq g(J)\right] \leq 12\text{e}^{-1.8b}$.

Proof of Lemma 5. Fix an arm i and attribute j. By the assumptions of the lemma, the stopping time τ_s is such that almost surely, each arm — and in particular arm i — has been pulled at least $s \geq s_0$ times at time τ_s. Define J to be the number of times that i has been pulled at time τ_s. For any $n \geq 1$, let k_n be the time step right after arm i has been pulled for the n^{th} time. Define $S_n := \frac{n\cdot(\hat{\mu}^{(j)}_{k_n,i} - \mu^{(j)}_i)}{2\sigma}$ to be the sum of all attribute-j noise components up to and including the n^{th} pull of arm i, renormalized to be a mean-zero sub-Gaussian$(1/2)$ random variable. The S_n define an unbiased half-subgaussian random walk, and we can therefore apply Lemma 11 to them and the stopping time J. Specifically, we set $b = \frac{x_J^2}{16\sigma^2}$, and obtain that

$$\text{Prob}\left[J < \infty \text{ and } S_J \geq \sqrt{0.6J \cdot \log(\log_{1.1}(J)+1) + \frac{Jx_J^2}{16\sigma^2}}\right] \leq 12\exp\left(\frac{-1.8Jx_J^2}{16\sigma^2}\right).$$

Applying Lemma 11 to $-S_n$ with the same choice of b, and taking a union bound over both cases, we obtain that

$$\text{Prob}\left[J < \infty \text{ and } |S_J| \geq \sqrt{0.6J \cdot \log(\log_{1.1}(J)+1) + \frac{Jx_J^2}{16\sigma^2}}\right] \leq 24\exp\left(\frac{-1.8Jx_J^2}{16\sigma^2}\right).$$

Because J will be finite with probability 1, we can drop the $J < \infty$ part of the event:

$$\begin{aligned}&\text{Prob}\left[S_J \geq \sqrt{0.6J \cdot \log(\log_{1.1}(J)+1) + \frac{J^2 x_J^2}{16\sigma^2}}\right]\\ =&\text{Prob}\left[J < \infty \text{ and } S_J \geq \sqrt{0.6J \cdot \log(\log_{1.1}(J)+1) + \frac{J^2 x_J^2}{16\sigma^2}}\right].\end{aligned}$$

In the high-probability case, we now apply the assumed inequality between x_J and x'_J, to obtain that

$$|S_J| \;\leq\; \sqrt{0.6J \cdot \log(\log_{1.1}(J)+1) + \frac{Jx_J^2}{16\sigma^2}} \;\leq\; \frac{Jx'_J}{2\sigma}.$$

Substituting the definition of S_J and canceling common terms, the inequality implies that $|\hat{\mu}^{(j)}_{k_J,i} - \mu^{(j)}_i| \leq x'_J$. The choice of J ensures that $\hat{\mu}^{(j)}_{k_J,i} = \hat{\mu}^{(j)}_{\tau_s,i}$, and we have thus shown that $|\hat{\mu}^{(j)}_{\tau_s,i} - \mu^{(j)}_i| \leq x'_s$. ■

Appendix D. Proof of Lemma 6

Lemma 6 *Let $x > 0$ be arbitrary. When $\mathcal{E}_t(x)$ happens, no agent $\boldsymbol{\theta}$ will pull a highly suboptimal arm, i.e., an arm i with $\boldsymbol{\theta} \cdot (\boldsymbol{\mu}_{B_{\boldsymbol{\theta}}} - \boldsymbol{\mu}_i) > 2Ddx$.*

Proof By definition, when $\mathcal{E}_t(x)$ happens, for all arms i and attributes j, all arm attribute estimates are accurate to within x, in the sense that $|\hat{\mu}^{(j)}_{t,i} - \mu^{(j)}_i| \leq x$.

Consider any agent type $\boldsymbol{\theta}$. Let $i \neq B_{\boldsymbol{\theta}}$ be any arm with much smaller true reward than the best arm: $\boldsymbol{\theta} \cdot (\boldsymbol{\mu}_{B_{\boldsymbol{\theta}}} - \boldsymbol{\mu}_i) > 2Ddx$. Because each coordinate of $\hat{\boldsymbol{\mu}}_{t,B_{\boldsymbol{\theta}}}$ and of $\hat{\boldsymbol{\mu}}_{t,i}$ is estimated accurately to within x, we get that $\boldsymbol{\theta} \cdot (\hat{\boldsymbol{\mu}}_{t,B_{\boldsymbol{\theta}}} - \boldsymbol{\mu}_{B_{\boldsymbol{\theta}}}) \geq -Ddx$ and $\boldsymbol{\theta} \cdot (\hat{\boldsymbol{\mu}}_{t,i} - \boldsymbol{\mu}_i) \leq Ddx$. Hence,

$$\begin{aligned}\boldsymbol{\theta} \cdot (\hat{\boldsymbol{\mu}}_{t,B_{\boldsymbol{\theta}}} - \hat{\boldsymbol{\mu}}_{t,i}) &= \boldsymbol{\theta} \cdot (\hat{\boldsymbol{\mu}}_{t,B_{\boldsymbol{\theta}}} - \boldsymbol{\mu}_{B_{\boldsymbol{\theta}}}) + \boldsymbol{\theta} \cdot (\boldsymbol{\mu}_{B_{\boldsymbol{\theta}}} - \boldsymbol{\mu}_i) + \boldsymbol{\theta} \cdot (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_{t,i})\\ &> -Ddx + 2Ddx - Ddx \;=\; 0,\end{aligned} \tag{2}$$

which means that the agent with type $\boldsymbol{\theta}$ will not pull arm i. ■

Appendix E. Proof of Lemma 7

Lemma 7 *Fix an arm i. Let $s \geq \exp(2/p)$, and let τ_s be the (random) time when arm i is pulled for the s^{th} time. Let $\hat{x} = \frac{1}{2Dd} \cdot \min(\hat{z}, \frac{p}{2L})$. Under $\mathcal{E}_{\tau_s}(\hat{x})$, this pull of arm i is not incentivized.*

Proof By Lemma 6, under the event $\mathcal{E}_{\tau_s}(\hat{x})$, all agent types $\boldsymbol{\theta}$ with $\boldsymbol{\theta} \cdot (\boldsymbol{\mu}_{B_{\boldsymbol{\theta}}} - \boldsymbol{\mu}_{B'_{\boldsymbol{\theta}}}) > 2Dd\hat{x}$ will pull their best arm $B_{\boldsymbol{\theta}}$. Notice that $2Dd\hat{x} = \min(\hat{z}, \frac{p}{2L})$

By Assumption 3, the measure of agents (across all arms) whose best and second-best arm differ in utility by less than $\min(\hat{z}, \frac{p}{2L})$ is at most $L \cdot \min(\hat{z}, \frac{p}{2L}) \leq \frac{p}{2}$. In particular, this bound holds for agents whose best arm is i. By Assumption 2, at least a measure p of agents has i as their best arm, and thus, at least a measure $\frac{p}{2}$ will myopically pull arm i. Because $1/\log(s) \leq \frac{p}{2}$ for $s \geq \exp(2/p)$, arm i is not payment-eligible at time τ_s. ■

Appendix F. Proof of Theorem 2

We restate the theorem here for convenience:

Theorem 2 *The expected total payment of Algorithm 1 is at most* $O\left(N^2 \cdot \mathrm{e}^{2/p}\right)$.

The high-level idea of the proof is motivated by Lemma 8, which shows that the expected *number* of payments is constant. Unfortunately, in contrast to the regret, there is no hard upper bound on the payment in any one round. If a draw of a particular arm comes out wildly inaccurate — which is an event of low but positive probability — then agents may require very large incentives to pull this arm again in the future (and correct the inaccurate estimate). The high payments are offset by the exceedingly low probability of having to incur them, but a rigorous analysis requires some care: if a high payment is required in one phase, this indicates a very inaccurate estimate, which may require multiple phases to correct. Hence, we need to handle dependency of payments across time steps and phases.

To reason about such estimation errors formally, we define *envelopes* of sample paths. A sample path ω captures all the random events affecting the algorithm, i.e., the random draws $\boldsymbol{\theta}_t$ of agents and the noise $\boldsymbol{\zeta}_t$ in the draws of the pulled arms, with an infinite time horizon.

With foresight, we define $g(s,\ell) := \frac{12\sigma\ell}{s^{2/5}}$. Let $\boldsymbol{\epsilon}_{t,i} = \hat{\boldsymbol{\mu}}_{t,i} - \boldsymbol{\mu}_i$ be the estimation error for the attribute vector $\boldsymbol{\mu}_i$ at time t, with components $\epsilon_{t,i}^{(j)}$. For any sample path ω, let $s(t,\omega)$ be the phase number that the algorithm is in at time t with the sample path ω. Define the sets

$$\begin{aligned}
\hat{\mathcal{L}}_\ell &= \{\omega \mid |\epsilon_{t,i}^{(j)}(\omega)| \le g(s(t,\omega),\ell) \text{ for all } i,j,t\},\\
\mathcal{L}_1 &= \hat{\mathcal{L}}_1,\\
\mathcal{L}_\ell &= \hat{\mathcal{L}}_\ell \setminus \hat{\mathcal{L}}_{\ell-1} \quad \text{for } \ell \ge 2.
\end{aligned}$$

We call $\mathcal{L}_\ell$ the ℓ^{th} envelope. In words, $\mathcal{L}_\ell$ consists of all sample paths such that at all times t, all coordinates of all arm estimation errors are bounded by $g(s,\ell)$, but for at least one time t, at least one coordinate of one arm estimation error is *not* bounded by $g(s,\ell-1)$. When ω is clear, we omit it in the notation for $\epsilon_{t,i}^{(j)}$, payments, etc. The importance of envelopes is that for small ℓ, the payments are tightly bounded, while for large ℓ, the cumulative probability of the sample paths in $\mathcal{L}_\ell$ is small. This is captured by the following two lemmas.

Lemma 12 *If $\omega \in \mathcal{L}_\ell$ and $s(t,\omega) = s$, then the payment in step t is upper-bounded by $\bar{c}(s,\ell) = R + 2Dd \cdot g(s,\ell)$.*

Proof The maximum payment is upper-bounded by the maximum perceived difference in value for any agent type and any two arms:

$$\begin{aligned}
\bar{c}(s,\ell) &\le \max_{\boldsymbol{\theta}} \left(\max_i \boldsymbol{\theta}\cdot\hat{\boldsymbol{\mu}}_{t,i} - \min_{i'} \boldsymbol{\theta}\cdot\hat{\boldsymbol{\mu}}_{t,i'} \right)\\
&= \max_{\boldsymbol{\theta}} \left(\max_i \boldsymbol{\theta}\cdot(\boldsymbol{\epsilon}_{t,i}+\boldsymbol{\mu}_i) - \min_{i'} \boldsymbol{\theta}\cdot(\boldsymbol{\epsilon}_{t,i'}+\boldsymbol{\mu}_{i'}) \right)\\
&\le \max_{\boldsymbol{\theta}} \left(\max_i \boldsymbol{\theta}\cdot\boldsymbol{\mu}_i - \min_{i'} \boldsymbol{\theta}\cdot\boldsymbol{\mu}_{i'} + \max_i \boldsymbol{\theta}\cdot\boldsymbol{\epsilon}_{t,i} - \min_{i'} \boldsymbol{\theta}\cdot\boldsymbol{\epsilon}_{t,i'} \right)\\
&\le R + 2Dd \cdot g(s,\ell).
\end{aligned}$$

The final inequality used the definition of the envelope. ■

Lemma 13 *For every $\ell \geq 2$, we have that* $\text{Prob}\left[\omega \in \mathcal{L}_\ell\right] \leq 24Nd\exp(-1.8(\ell-1)^2)$.

Proof For ω to be in $\mathcal{L}_\ell$, by definition, for at least one time t, at least one coordinate of at least one arm's estimation error must exceed $g(s(t,\omega),\ell-1)$. For now, fix an arm i and coordinate j.

Define $x_s := \frac{4\sigma(\ell-1)}{s^{1/2}}$ and $x'_s := g(s,\ell-1) = \frac{12\sigma(\ell-1)}{s^{2/5}}$.

Recall that $m_{t,i}(\omega) \geq s(t,\omega)$ is the number of times that arm i has been pulled by time t under ω. Let the random stopping time $\tau_{i,j}$ be the first value of $m_{t,i}$ (i.e., the first pull of arm i) such that the estimation error of attribute j of arm i exceeds $g(m_{t,i},\ell-1)$. $\tau_{i,j} = \infty$ means that the estimation error never exceeds $g(m_{t,i},\ell-1)$. We next verify that x_s, x'_s as we defined them satisfy the assumption

$$\sqrt{0.6n\cdot\log(\log_{1.1}(n)+1)+\frac{nx_n^2}{16\sigma^2}} \;\leq\; \frac{nx'_n}{2\sigma}$$

of Lemma 5, for all $n \geq 1$. First, we show that $\sqrt{0.6n\log(\log_{1.1}(n)+1)} \leq 5n^{3/5}$ for all $n \geq 1$. By squaring the inequality and canceling out a factor n, the statement is equivalent to showing that $25n^{1/5} \geq \log(\log_{1.1}(n)+1)$. Indeed, a stronger statement holds, namely, that $25n^{1/5} \geq \log_{1.1}(n)+1$. To see this, notice that the derivative of the left-hand side is always strictly larger than the derivative of the right-hand side, so the difference between the sides is minimized at $n=1$, where it is positive. Using this inequality and subadditivity of $\sqrt{\cdot}$, we can bound

$$\sqrt{0.6n\cdot\log(\log_{1.1}(n)+1)+\frac{nx_n^2}{16\sigma^2}} \;\leq\; 5n^{3/5}+\frac{\sqrt{n}x_n}{4\sigma} \;=\; 5n^{3/5}+(\ell-1) \;\leq\; \frac{nx'_n}{2\sigma}.$$

Applying Lemma 5 to $\tau_{i,j}$, we conclude that the probability that the error in coordinate j of arm i exceeds $x'_s = g(s,\ell-1)$ at time $\tau_{i,j}$ (and hence at any time, by definition of $\tau_{i,j}$) is at most $24\exp\left(-\frac{1.8sx_s^2}{16\sigma^2}\right) = 24\exp(-1.8(\ell-1)^2)$. Now, taking a union bound over all arms i and coordinates j completes the proof. ■

Next, we show that for any sample path ω in the envelope $\mathcal{L}_\ell$, we can bound the total number of payments made in terms of ℓ. Define $h(\ell) := \max\left(\exp(\frac{2}{p}), \left(\frac{24\sigma\ell Dd}{\hat{z}}\right)^{5/2}, \left(\frac{48\sigma\ell LDd}{p}\right)^{5/2}\right)$.

Lemma 14 *Let ω be a sample path in $\mathcal{L}_\ell$. Then, under ω, the algorithm makes payments at most for the first $h(\ell)$ phases.*

Proof The proof is similar to that of Lemma 8. Fix a sample path $\omega \in \mathcal{L}_\ell$. Consider a phase $s > h(\ell)$ and a time t in phase s. Because $\omega \in \mathcal{L}_\ell$, in particular, all estimation errors are bounded at time t, in the sense that the event $\mathcal{E}_t(g(s,\ell))$ happens. Because $s \geq h(\ell)$, we get that $g(s,\ell) \leq \min(\frac{\hat{z}}{2Dd}, \frac{p}{4DdL})$ as well as $s \geq \exp(2/p)$. Therefore, we can apply Lemma 7 and conclude that the pull at time t was not incentivized. ■

Proof of Theorem 2. We can write the total expected payment as

$$\mathbb{E}\left[\sum_{t=1}^{\infty} c_t\right] = \sum_{\ell=1}^{\infty}\sum_{\omega\in\mathcal{L}_\ell} \text{Prob}\left[\omega\right]\cdot\sum_{t=1}^{\infty} c_t(\omega) \;=\; \sum_{\ell=1}^{\infty}\sum_{\omega\in\mathcal{L}_\ell} \text{Prob}\left[\omega\right]\cdot\sum_{s=1}^{\infty}\sum_{t:s(t,\omega)=s} c_t(\omega).$$

By Lemma 14, the payments are 0 for $s > h(\ell)$, and by Lemma 12, when $\omega \in \mathcal{L}_\ell$ and $s(t,\omega) = s$, we can bound $c_t(\omega) \le R + 2Dd\cdot g(s,\ell)$. Furthermore, in any one phase, because each arm is incentivized at most once, there are at most N payments total. Substituting these bounds, we obtain that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{\infty} c_t\right] &\le \sum_{\ell=1}^{\infty}\sum_{\omega\in\mathcal{L}_\ell} \text{Prob}\left[\omega\right]\cdot\sum_{s=1}^{h(\ell)} N\cdot(R+2Dd\cdot g(s,\ell))\\ &= N\cdot\sum_{\ell=1}^{\infty}\text{Prob}\left[\omega\in\mathcal{L}_\ell\right]\cdot\sum_{s=1}^{h(\ell)}\left(R+\frac{24Dd\sigma\ell}{s^{2/5}}\right).\end{aligned}$$

We now lower-bound $s^{2/5} \ge 1$, split off the term for $\ell = 1$ and bound $\text{Prob}\left[\omega\in\mathcal{L}_1\right] \le 1$, and apply Lemma 13 to the remaining $\text{Prob}\left[\omega\in\mathcal{L}_\ell\right]$ terms, to bound

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{\infty} c_t\right] &\le N\cdot\sum_{s=1}^{h(1)}(R+24Dd\sigma) + N\cdot\sum_{\ell=2}^{\infty}24Nd\exp(-1.8(\ell-1)^2)\cdot\sum_{s=1}^{h(\ell)}(R+24Dd\sigma\ell)\\ &\le N\cdot h(1)\cdot(R+24Dd\sigma) + 24N^2d\cdot\sum_{\ell=2}^{\infty}\exp(-1.8(\ell-1)^2)\cdot h(\ell)\cdot(R+24Dd\sigma\ell).\end{aligned}$$

Because $h(\ell)$ and $(R+24Dd\sigma\ell)$ grow polynomially in ℓ, whereas $\exp(-1.8(\ell-1)^2)$ decreases exponentially in ℓ, the sum is dominated by its first term, and the overall expected payment is bounded by

$$\begin{aligned}&O\left(N^2d\cdot(R+Dd\sigma)\cdot h(1)\right)\\ =&O\left(N^2d\cdot(R+Dd\sigma)\cdot\max\left(\exp\left(\frac{2}{p}\right),\left(\frac{\sigma Dd}{\hat{z}}\right)^{5/2},\left(\frac{\sigma LDd}{p}\right)^{5/2}\right)\right)\\ =&O\left(N^2\cdot\exp\left(\frac{2}{p}\right)\right).\end{aligned}$$

■

Appendix G. Proof Sketch of Theorem 9

Theorem 9 *Under Assumption 4, the Discrete-Preference Algorithm has expected payment bounded above by $O\left(N^2/p\right)$ and expected regret bounded above by $O(N/p)$.*

Proof The proof of the expected payment bound is the same as the proof of Theorem 2, except that we now define $h(\ell) := \max\left(\frac{2}{p}, \left(\frac{24\sigma\ell Dd}{\hat{z}}\right)^{5/2}\right)$.

To prove the bound on the expected regret, one can first prove tightened versions of Lemmas 7 and 8, which replace the $\exp(2/p)$ term with $2/p$ and use $L = 0$. In return, the length of phase s is now bounded only by Ns instead of $N\log(s)$, and the expected number of times steps in which a payment is made is bounded by $O(N/p + N^2)$. Substituting these changes into the proof of Theorem 3, we obtain the bound $O(R(N/p + N^2) + NR + NR) = O(N/p)$ since $p \leq 1/N$. ■

Appendix H. Proof Sketch of Theorem 10

Theorem 10 *Under Assumption 5, the Known-p Algorithm has expected payment bounded above by* $O(N^2 \cdot \max(1, (L/p)^{5/2}))$ *and expected regret bounded above by* $O\left(\frac{N^2}{p^2} + \frac{NL\log^3(T)}{p}\right)$.

Proof The proof of the expected payment bound follows Lemma 2, except we define $h(\ell)$ without including the $\exp(2/p)$ term, instead using $h(\ell) := \max\left(\left(\frac{24\sigma\ell Dd}{\hat{z}}\right)^{5/2}, \left(\frac{48\sigma\ell LDd}{p}\right)^{5/2}\right)$. The resulting bound on the expected payment is

$$O\left(N^2 d \cdot (R + Dd\sigma) \cdot \max\left(\left(\frac{\sigma Dd}{\hat{z}}\right)^{5/2}, \left(\frac{\sigma LDd}{p}\right)^{5/2}\right)\right) = O\left(N^2 \max(1, (L/p)^{5/2})\right).$$

The proof of the expected regret bound first establishes tightened versions of Lemmas 7 and 8, proving the following upper bound on the number of time steps in which a payment is made:

$$O\left(\frac{N^2L^3D^3d^3\sigma^3}{p^2} + \frac{N^2d^3\sigma^2D^2}{\hat{z}^2} + \frac{ND^3d^3\sigma^3}{\hat{z}^3}\right) = O\left(\frac{N^2}{p^2}\right).$$

The proof of this result follows that of Lemma 8, but the less aggressive incentivization allows us to define $s_1 = \max(2, \frac{30\sigma^3}{x^3})$ since $\frac{1}{\log(s+s_0)} \leq \frac{p}{2}$ is true for all s.

Using this tighter bound on the number of incentivizations, and the fact that phases now last at most $N\log(s+s_0)$ steps in expectation, we can bound the regret in Equation (1) by $O\left(\frac{ND^2d^2\sigma^2L\cdot(\log(T))^3}{p}\right)$.
■